Paper 13 PAMS (15) 13

# Population And Migration Statistics (PAMS) Committee (Scotland)

## Investigating quality measures for mid-year estimates

**PAMS members' views on the usefulness of the types of measures outlined in this paper are welcomed. We think that they will be useful for the quality assurance stage of the mid-year estimates production process but thoughts on their usefulness for users of our statistics are also welcomed.**

### Contents

### List of Tables

## Main Points

The main points in this report are:

**We are aiming to publish more information about the quality of mid-year population estimates…**

- The mid-year population methodology follows the internationally recognised cohort-component method.
- Accuracy of the mid-year population estimates are only quantified for headline council area population estimates in a census year.
- This is particularly important for users interested in year-on-year changes and for statisticians to identify statistically significant population change for council areas. For National Records of Scotland (NRS), robust confidence intervals would help identify parts of the methodology to develop and would give users more information about their quality.

**… but it is difficult.**

- Each component of the mid-year population estimation methodology is susceptible to different kinds of uncertainty that require different statistical methodologies.
- The uncertainty for international migration and the census base components of the methodology are relatively straightforward to estimate.
- The uncertainty for internal migration is more difficult to quantify since error can only be inferred in a census year. Some work has been carried by Office for National Statistics (ONS) who have proposed a methodology which offers a basis for estimating error rates.
- Estimates of special populations such as prison and armed forces populations are susceptible to respondent errors. Uncertainty for these populations are difficult to measure, but will likely affect a small number of council areas.

**This paper provides provisional measures using two methodologies, but the validation of these results are difficult.**

- An error component methodology is proposed, based on similar work carried out by ONS. This methodology estimates the error for each component of the mid-year population estimates (MYE) methodology. These provide an indication of confidence for a point estimate and can be used to infer the statistical significance of council area population change.
- A stability approach is also proposed, that estimates the level of random variation of a time-series. This methodology overcomes some of the shortcomings of the error component methodology but cannot estimate the confidence intervals of point estimates.
- Both approaches have drawbacks and benefits. In detecting significant population change between 2011-2012, the results show that different inferences would be made for nine council areas from using both methodologies.

## 1. Purpose of Report

This report investigates the feasibility of producing a measure of uncertainty for MYE that can accompany their publication.

In so doing, the report identifies the two main potential sources of error in the mid-year estimates and presents two possible methods to calculate them.

This report addresses the questions:

- What are the main sources of error in the mid-year population estimates?
- Can these sources of error be quantified meaningfully to produce some indication of confidence of a mid-year population estimate?

## 2. Why are confidence intervals important?

A census currently provides the most accurate population estimates in Scotland. The census year provides the only time where some indication of error can be calculated for a Scottish population estimate.

For many reasons, including costs and respondent fatigue, it is only possible to conduct a census every ten years. To estimate the population in-between censuses, the inter-census period, MYEs are calculated using the cohort component method. This involves ageing the existing stock of the population, and estimating net flows of natural population change (births and deaths) and migration (both internal migration and international migration flows). Since there are a variety of different component methodologies contributing to a MYE, there is difficulty in estimating confidence intervals.

It should be expected that as the MYEs are rolled forward, the further a MYE gets from a census year the more uncertainty there will be in the point estimate. The main mechanism used to estimate error in these rolled forward estimates is by comparing the rolled forward MYE estimates in a census year with the results from the census in that year. This should identify any in-built bias within the MYE methodology and facilitates development work to improve the methodology for the next round of MYEs.

The production of confidence intervals for population estimates becomes particularly important for interpreting year-on-year population change for council areas further from a census year. Within NRS’ population projection methodology, for example, flow data is pooled across five years to help eliminate the possible effects of random variation within annual migration flows. A robust confidence interval should allow users to better understand if moderate annual net population change is likely to be statistically significant and to better inform NRS statisticians when and how data should be pooled.

## 3. Structure of the report

This report first describes the sources of uncertainty in the MYEs.

The report then describes two methodologies that could be used in estimating uncertainty of the MYEs.

Finally, the application of these methodologies is applied to 2012 data before discussing potential methodological issues if they are to be used in the future.

## 4. Sources of uncertainty in MYEs

### 4.1 The census

While the census provides very accurate data for national and council area headline population estimates, there are still some sources of error in the estimation of non-response. For council areas and age groups, the reported confidence intervals in the census is an indication of the initial level of uncertainty at the start of the inter-census period. The absolute level of uncertainty in a census estimate should not diminish throughout the inter-census period, and can be considered 'peak-year' accuracy.

### 4.2 International migration

International migration is primarily estimated through the International Passenger Survey (IPS) for Scotland. This is used as the basis to allocate individuals to council areas using  administrative data. There are, then, two sources of potential error from international migration estimates for total council area populations; sampling error within the IPS and measurement errors in the administrative data that is used in allocation. Each year the uncertainty in flows (both inward and outward) result in increasing uncertainty in the rolled forward population stock. This component will have a cumulative impact and contributes to increasing uncertainty in the population base every year.

### 4.3 Internal migration

Internal migration flows (both inward and outward) also have a cumulative impact on uncertainty. Internal migration flows are almost wholly estimated using administrative data. The source of error in this component is measurement errors within administrative data. This is caused by individuals moving into or out of an area but not registering with a General Practitioner (GP), or delaying registration with a GP.

### 4.4 Other sources of error

The other main component of the MYE methodology is natural change (the difference between the number of births and deaths). Measurement errors in the birth and death register would contribute to uncertainty in the same way as internal migration. It is generally assumed that errors on these lists are negligible. If this assumption is valid, positive net natural change would 'reduce' overall uncertainty in council area population estimates through increasing the denominator. Similarly, negative net natural change would 'increase' uncertainty by decreasing the denominator.

The final components, estimates of the prison and armed forces populations, will also contribute to uncertainty. This is particularly important in specific council areas with large numbers of these populations. These components generally rely on returns from individual establishments. Errors in these returns, response error, may also contribute to uncertainty. It is unclear whether these errors will have a cumulative impact. It might be safe to assume that inter-census estimates for these components are independent of one another (i.e. bias reported in a force return one year is not replicated the next for the same establishment). If so, this would create a certain level of error in population stocks for each year. If significant, this problem would likely affect inferences for year-on-year changes for any given single year. Methodological issues may create bias, but it should be assumed that post-census development work identifies and corrects for these issues.

## 5 . Current methodologies to estimate sources of uncertainty

### 5.1 Quality indicators

The results contained in the census-MYE reconciliation report and published census confidence intervals are the only quantified uncertainty measures that can be used for validation of derived confidence intervals. Since inter-census confidence intervals are not available, there is some difficulty in determining credible lower and upper bounds. Since the work is at a preliminary stage, over- or under- estimates of uncertainty can be more easily attributed to methodological assumptions rather than to actual uncertainty within the data.

This being so, it still might be possible to provide users with some indication of the likely sources of errors. Identifying hard to count populations, their relative importance in the MYE methodology is one way this can be achieved. Such an exercise would not necessarily provide any indication of confidence, but provide an indication of which local authorities have disproportionately larger numbers of these groups. But while this would provide indication of how hard a population is to count, these indicators would not facilitate testing the statistical significance of annual population change.

### 5.2 Estimating the Census error component

The census error component for Scotland and its council areas is estimated using a bootstrap methodology and have been provided for the 2011 census.

### 5.3 Estimating the International migration error component

Sampling error can be estimated directly for the IPS based on its sample size. It is possible to allocate the upper and lower bounds of the national confidence interval directly to council areas to derive confidence limits for this component.

Errors in the allocation methodology would likely take the form of measurement errors, and are difficult to observe. Further development work to understand the contribution of the allocation methodology to uncertainty would have to be undertaken.

The key assumption within this component is that IPS sampling error is a unique national problem, and that the variance can be allocated to local authorities.

5.4 Estimating the Internal migration error component

There is no obvious method to quantify error in administrative datasets. Only in a census year, where it can be assumed that the difference between flows reported in administrative datasets and those reported in the census, can some indication of potential error be estimated.

ONS proposed using a methodology based on the distribution of residuals from a non-linear regression model but a more simple derivation using a component method is proposed here.

A scaling factor is estimated for each ‘gad’ component, where

$$l_{gad} = \frac{c_{gad}}{m_{gad}}$$

$l$ is the scaling factor for an age-sex group, $a$, in a council area, $g$, and for both internal migration inflow and outflow, $d$, using the census estimate of flow, $c$, and administrative data, $m$. With no clustering or stratification, the distribution of scaling factors are assumed to have an approximately $N(1,\sigma^2)$ sampling distribution across all council areas for each ‘ad’ component.

A bootstrapping methodology resamples observations from this sampling distribution. For each replicate, the product of a randomly selected $l_{gad}$ and estimated internal migration flow from administrative sources provides a point on the estimated error distribution for a given council area’s ‘ad’ component. From these components in each replicate, the simulated inflow and outflow error distributions can be derived. Central to this methodology is, therefore, an homogeneity assumption and that error distributions can be observed from observations across council areas. One potential issue with this methodology is assuming that variance over balance should be used and not the uncorrelated sum of inflow and outflow variances. The latter will substantially increase the level of reported variation.

5.5 Estimating other error components

Respondent error in the prison and armed forces population, are more difficult to quantify. The causes of error can range from questionnaire design to seasonal effects. These error components have relatively a small impact on the Scottish population, but could have disproportionately large local impacts.

5.6 Estimating random variation from a time-series

Scotland produces population projections every two years. These projections use a five year average to stabilise migrant flows to generate trends. This is because there will be some random variation in any given MYE.

Using this as a basis to estimate uncertainty can overcome the major methodological issues associated with inferring an error distribution from observations across heterogeneous council areas and the error associated with allocating migrants to council areas. A more coherent approach, then, might be to estimate the stability of population change. This approach is based on the assumption that a single council area’s population dynamics do not change drastically year-on-year, and only when they do will there be uncertainties in population estimation.

Current demographic modelling assumptions in public statistics make three core assumptions about the population; there will be random variation year-on-year that can to be stabilised and that MYEs will follow a stochastic process through time but still likely follows a meaningful trend (i.e., the MYE mean over time is not constant). Using these assumptions, transforming a time-series to a stationary process is possible in order to derive variance over time. The variance of this distribution would imply a level of random variation using observations from the same council area. While many methods are available to meet this goal, the most obvious is de-trending for a short time-series. This methodology is the most consistent with general NRS assumptions used in generating population projections, and so is proposed here.

Since mid-year population estimates are only available on an annual basis, and so there are limited numbers of observations, the ‘true’ trend, and stability, of the series may be difficult to capture. Breaks in the time-series can affect trends with low numbers of observations, so detecting genuine changes in population dynamics from random variation would have to be based on assumptions.

## 6 . Results

Results are presented in this paper for 2012, using both the error component and stability methodologies described in section 5.2 to 5.6.

### 6.1 Stability of MYEs

A major problem predicting possible changes in the rate of growth of a council area’s population is that dynamics can change over a short period of time (10 years). This type of phenomena is not uncommon in population change for Scotland’s council areas. A standard method to detect and account for these problems can be found in econometric time-series regression analysis by detecting structural breaks over time and incorporating them into a statistical model.

These types of solutions are not feasible without a very long time-series, and so could not be produced every year. Instead, a quadratic regression model is proposed to estimate the trend. Such a functional form would allow one or two ‘breaks’ (i.e., a point in which population dynamics have changed) in population growth rates in a ten-year time-series. Further breaks in the time-series would be considered to increase instability and are not attributed to be model misspecification. This assumption is driven by statistical concerns, the number of observation points in a time-series would prohibit the use of a higher order polynomial function.

Table 1 shows the results of applying this methodology to MYEs for 2003-2010 based on the last census (2001) and uses 2011 and 2012 data rolled forward from the 2011 census. The benefit of using these data is a likely ‘break’ in the series for most council areas resulting from the census 2011. Measuring predictability over such a break provides indication consistency over time for each council areas. The table also shows the population change between 2011 and 2012. It suggests that, using the rebased MYEs, twelve council areas (37 per cent) witnessed no significant change in the population (P(α)=0.05). When including the non-rebased 2002-2010 data, the time-series becomes more unstable (by including a further break in the trend), so much so that as much as 80 per cent of council areas would have reported no statistically significant population change.

**Table 1: Random variation around quadratic trend**

| | Non-Rebased<br>95% Confidence Interval | Rebased<br>95% Confidence Interval | 2011-2012 Change |
|---|---|---|---|
| Aberdeen City | 0.6% | 0.6% | 1.1% |
| Aberdeenshire | 0.8% | 0.2% | 0.7% |
| Angus | 1.9% | 0.2% | 0.0% |
| Argyll & Bute | 0.7% | 0.9% | -2.3% |
| Clackmannanshire | 0.9% | 0.9% | -0.4% |
| Dumfries & Galloway | 0.9% | 0.2% | -0.4% |
| Dundee City | 0.5% | 0.3% | 0.4% |
| East Ayrshire | 0.6% | 0.3% | 0.0% |
| East Dunbartonshire | 0.4% | 0.3% | 0.8% |
| East Lothian | 0.8% | 0.9% | 0.9% |
| East Renfrewshire | 0.5% | 0.2% | 0.2% |
| Edinburgh, City of | 1.1% | 0.4% | 1.0% |
| Eilean Siar | 2.4% | 0.4% | -0.5% |
| Falkirk | 0.7% | 0.2% | 0.4% |
| Fife | 0.1% | 0.2% | 0.3% |
| Glasgow, City | 0.4% | 0.6% | 0.3% |
| Highland | 1.6% | 0.3% | 0.1% |
| Inverclyde | 0.9% | 0.3% | -0.7% |
| Midlothian | 0.7% | 0.5% | 0.9% |
| Moray | 2.2% | 0.8% | -0.6% |
| North Ayrshire | 0.8% | 0.3% | -0.4% |
| North Lanarkshire | 1.2% | 0.3% | 0.0% |
| Orkney Islands | 2.3% | 0.3% | 0.5% |
| Perth & Kinross | 0.9% | 0.5% | 0.6% |
| Renfrewshire | 0.9% | 0.4% | -0.2% |
| Scottish Borders | 0.5% | 0.5% | -0.1% |
| Shetland Islands | 1.3% | 0.8% | -0.1% |
| South Ayrshire | 0.6% | 0.2% | -0.1% |
| South Lanarkshire | 0.2% | 0.1% | 0.1% |
| Stirling | 0.5% | 0.5% | 0.8% |
| West Dunbartonshire | 0.2% | 0.3% | -0.3% |
| West Lothian | 0.5% | 0.4% | 0.4% |

For simplicity, headline MYEs have been de-trended. A variance could also be estimated for each source of uncertainty. Using the sum of variances, a different measure might be derived, which will increase the width of the confidence intervals reported in table 1.

## 6.2 Estimating separate components

The error component methodology more explicitly estimates confidence intervals around the MYE point estimate for a given year. Statistically significant population change can be detected through identifying overlapping 84 per cent confidence intervals between years (P(α)=0.05). Table 2 shows the results of applying this methodology to data in 2012. For simplicity, it is assumed that the census base approximates to the error in the 2011 MYEs.

The methodology results in a relatively slow degradation of the confidence intervals through time, but this may be due to methodological issues. Despite this, the significance testing here suggests that it is very difficult to detect any significant annual change for any council area (Argyll & Bute the only council area which would allow the null hypothesis of no significant change to be rejected).

**Table 2: Detecting significant annual change using point estimates**

| | Census 2011 CI 95% | MYE 2012 C.I. 95% | Change from Census2011 - MYE2012 | Significant Change 2011-2012 P = 0.05 |
|---|---|---|---|---|
| Aberdeen City | 2.6% | 2.7% | 1.1% | FALSE |
| Aberdeenshire | 1.2% | 1.2% | 0.7% | FALSE |
| Angus | 1.3% | 1.3% | 0.0% | FALSE |
| Argyll & Bute | 1.4% | 1.4% | -2.3% | TRUE |
| Clackmannanshire | 1.0% | 1.0% | -0.4% | FALSE |
| Dumfries and Galloway | 0.8% | 0.8% | -0.4% | FALSE |
| Dundee City | 2.2% | 2.3% | 0.4% | FALSE |
| East Ayrshire | 0.9% | 1.0% | 0.0% | FALSE |
| East Dunbartonshire | 1.4% | 1.5% | 0.8% | FALSE |
| East Lothian | 1.1% | 1.1% | 0.9% | FALSE |
| East Renfrewshire | 1.6% | 1.6% | 0.2% | FALSE |
| Edinburgh, City of | 1.8% | 2.0% | 1.0% | FALSE |
| Eilean Siar | 1.5% | 1.5% | -0.5% | FALSE |
| Falkirk | 1.0% | 1.1% | 0.4% | FALSE |
| Fife | 1.1% | 1.2% | 0.3% | FALSE |
| Glasgow City | 1.8% | 2.0% | 0.3% | FALSE |
| Highland | 1.5% | 1.5% | 0.1% | FALSE |
| Inverclyde | 1.7% | 1.7% | -0.7% | FALSE |
| Midlothian | 1.6% | 1.6% | 0.9% | FALSE |
| Moray | 1.5% | 1.6% | -0.6% | FALSE |
| North Ayrshire | 1.0% | 1.0% | -0.4% | FALSE |
| North Lanarkshire | 1.4% | 1.4% | 0.0% | FALSE |
| Orkney Islands | 2.3% | 2.4% | 0.5% | FALSE |
| Perth and Kinross | 1.2% | 1.3% | 0.6% | FALSE |
| Renfrewshire | 1.5% | 1.6% | -0.2% | FALSE |
| Scottish Borders | 1.0% | 1.1% | -0.1% | FALSE |
| Shetland Islands | 1.3% | 1.3% | -0.1% | FALSE |
| South Ayrshire | 0.8% | 0.9% | -0.1% | FALSE |
| South Lanarkshire | 1.2% | 1.2% | 0.1% | FALSE |
| Stirling | 1.3% | 1.6% | 0.8% | FALSE |
| West Dunbartonshire | 1.6% | 1.6% | -0.3% | FALSE |
| West Lothian | 1.4% | 1.4% | 0.4% | FALSE |

However, detecting significant annual change in this way is problematic since the census base is not an indication of random error each year, but an indication of possible bias in the MYEs from the census year. It is not possible to observe this effect over time, but it might be possible to make the assumption that this residual bias might be less relevant in testing the significance of annual change. In other words, the MYEs are primarily interested in estimating flows into and out of a council area, and are, therefore, only interested in the degree of confidence around net changes.

Subject to these assumptions[1], removing the census base error component from analysis might provide a 'flow confidence interval'; the degree of uncertainty around the flows into and out of a council area. Table 3 shows the results of producing these confidence intervals to infer significance. Assuming 2011 has no error component[2], just under half (47 per cent) of the council areas' MYEs in 2012 did not experience statistically significant population change.

**Table 3: Detecting significant annual change using 'flow confidence intervals'**

| | MYE 2012 Flow C.I. 95% | Change from Census2011 - MYE2012 |
|---|---|---|
| Aberdeen City | 0.8% | 1.1% |
| Aberdeenshire | 0.3% | 0.7% |
| Angus | 0.2% | 0.0% |
| Argyll & Bute | 0.4% | -2.3% |
| Clackmannanshire | 0.3% | -0.4% |
| Dumfries and Galloway | 0.3% | -0.4% |
| Dundee City | 0.8% | 0.4% |
| East Ayrshire | 0.2% | 0.0% |
| East Dunbartonshire | 0.3% | 0.8% |
| East Lothian | 0.3% | 0.9% |
| East Renfrewshire | 0.3% | 0.2% |
| Edinburgh, City of | 1.0% | 1.0% |
| Eilean Siar | 0.3% | -0.5% |
| Falkirk | 0.3% | 0.4% |
| Fife | 0.5% | 0.3% |
| Glasgow City | 0.7% | 0.3% |
| Highland | 0.3% | 0.1% |
| Inverclyde | 0.2% | -0.7% |
| Midlothian | 0.3% | 0.9% |
| Moray | 0.4% | -0.6% |
| North Ayrshire | 0.3% | -0.4% |
| North Lanarkshire | 0.2% | 0.0% |
| Orkney Islands | 0.3% | 0.5% |
| Perth and Kinross | 0.5% | 0.6% |
| Renfrewshire | 0.3% | -0.2% |
| Scottish Borders | 0.4% | -0.1% |
| Shetland Islands | 0.3% | -0.1% |
| South Ayrshire | 0.3% | -0.1% |
| South Lanarkshire | 0.2% | 0.1% |
| Stirling | 0.9% | 0.8% |
| West Dunbartonshire | 0.3% | -0.3% |
| West Lothian | 0.3% | 0.4% |

**Footnotes**

[1] The assumption is that the census error component does not directly affect the error reported in internal and international migration flows and is, therefore, not relevant when estimating uncertainty in inter-census population change. The census error component will have some impact on uncertainty for subsequent MYEs (e.g., in allocation) so whether this simplification can be made would have to be investigated.

[2] For the 2011 MYE this might be a safe assumption, for future tests of significance a similar test to table 2 would have to be undertaken. If we assumed that 2011 flows have similar levels of uncertainty as 2012 flows, the results suggest that there would be difficulties in determining significant change for just under three-quarters of the council areas.

These results, in some way, offer an opportunity to validate results of two methodologies. Comparing the results reported in table 1 with those reported in table 3 show that the same inference would be made for twenty-three council areas (71 per cent). However, different inferences would be drawn for eight council areas (29 per cent). This is most likely due to the fact that the level of reported error in internal migration scales with the absolute level of net internal migrants, so that Edinburgh, for example, reports relatively larger error rates using the component method. Both methodologies rely on differing assumptions so it is difficult to determine which is the most suitable at this stage.

## 7 . Recommendations

The methodologies outlined here offer two ways to detect significant annual population change for MYEs. Likewise, it offers a methodology that can be used to derive confidence intervals for point estimates for any given year.

The error component methodology is the most ambitious and offers confidence intervals around the point estimate as well as the opportunity to infer statistical significance for annual change. The results of this method suggest that it is relatively difficult to determine if annual population changes in council areas are statistically significant. Four main areas will need to be investigated, or noted, if this methodology is pursued in the long-term;

- The assumptions used to estimate error in the internal migration component. This is currently based on the variance of error rates across heterogeneous council areas and an error distribution that will simply be rescaled throughout the inter-census period.
- The need to develop some indication of variation in the allocation of international migrants to council areas.
- The way in which the census base ought to be incorporated into statistical testing of net annual changes.
- Ways to incorporate an error component of special populations into the methodology.

The stability approach overcomes the problem of using heterogeneous council areas to quantify random variation and, in theory, incorporates the special populations' error component. The methodology offers a way to monitor the significance of net annual change based on the assumptions used to develop population projections. This methodology is not able to provide confidence intervals of point estimates. The results of this methodology suggest that it is difficult to determine significant population change for just over a third of council areas in 2012. Three main areas will need to be investigated, or noted, if this methodology is pursued in the long-term;

- The types of trends that should be assumed in de-trending the MYE time-series over time. This is linked to the frequency of flow data used in the MYE methodology.
- The way in which the census base error component is dealt with in analysis and inference and whether, more generally, the variance of each component should be measured individually.

- The extent to which other forms of bias and causal factors may explain observed variation of a council area’s time-series, and how they might be incorporated into a statistical model.

Population and Migration Statistics
29 April 2015